Vine Copula Structure Learning via Monte Carlo Tree Search Supplementary Materials

1 From Vine to Multivariate Distribution

Detailed illustration of the process of obtaining a multivariate distribution from vine based on the example in Figure 1 in the main text is given below. For tree 1, edges can be assigned bivariate distributions $F_{12}, F_{13}, F_{24}, F_{25}$ given the univariate marginal distributions F_1, F_2, F_3, F_4, F_5.For tree 2, edges can be assigned the conditional distributions $F_{23|1}, F_{14|2}, F_{45|2}$; for example, $F_{23|1}$ summarizes the conditional dependence of $F_{2|1}, F_{3|1}$ where $F_{2|1}, F_{3|1}$ can be obtained from F_{12}, F_{13} in tree 1 respectively. The combination of $F_{23|1}, F_{12}, F_{13}$ yields the trivariate distribution F_{123}. For tree 3, edges can be assigned the conditional distributions $F_{34|12}, F_{15|24}$; for example $F_{34|12}$ summarizes the conditional dependence of $F_{3|12}, F_{4|12}$, which can be obtained from F_{123}, F_{124}. As mentioned above, F_{123}, F_{124} can be achieved from the combining conditional distributions in trees 1 and 2.

There are bivariate distributions on the edges in trees 1 to $d-1$ of the vine. If the bivariate distributions on the edges are all bivariate Gaussian, each edge can be characterized by a correlation parameter ρ, which can be interpreted as a partial correlation for trees 2 to $d-1$. For the above example, one could consider that the edges have been assigned the quantities $\rho_{12}, \rho_{13}, \rho_{24}, \rho_{25}, \rho_{23;1}, \rho_{24;1}, \rho_{45;2}, \rho_{34;12}, \rho_{15;24}$; here the semicolon in the subscript is common for the partial correlation. For example, $\rho_{15;24}$ summarizes the conditional correlation of variables 1 and 5 given variables 2 and 4. Partial correlations can be calculated by inverting the principal submatrix of a correlation matrix. Specifically, consider a partial correlation $\rho_{a,b;S}$ where S is a set of variables and $\{a,b\} \cap S = \emptyset$. Let $\mathbf{\Sigma}$ be the correlation matrix of $\{a,b\} \cup S$. If we define $\mathbf{\Omega} = (\omega_{ij}) = \mathbf{\Sigma}^{-1}$, we have $\rho_{a,b;S} = -\omega_{ab}/\sqrt{\omega_{aa}\omega_{bb}}$.

The representation of a multivariate Gaussian distribution through a vine is an alternative parametrization of the correlation matrix that avoids the positive definiteness constraint of a correlation matrix. From Kurowicka and Cooke (2003) and Kurowicka and Cooke (2006), the correlations and partial correlations assigned to any vine are algebraically independent. and the determinant of the correlation matrix is $\log\det(\mathbf{R}) = \prod_e (1-\rho_e^2)$ for any vine with $\{\rho_e\}$ being the set of correlations and partial correlations on the edges of the vine. Moreover, it is this parametrization of multivariate Gaussian that can extend to multivariate non-Gaussian by using bivariate copulas on the edges of the vine to get what is called the vine copula or pair-copula construction.

Multivariate data are seldom well summarized by the multivariate Gaussian distribution, but the multivariate Gaussian may be adequate as a first order model if the variables are monotonically related to each other. One approach to developing a parsimonious copula for high-dimensional non-Gaussian data is to (a) find a parsimonious truncated partial correlation vine for the matrix of normal scores (where variables have each been converted to standard normal via probability integral transforms), and (b) replace edges of the vine with bivariate copulas that can have tail behavior different from Gaussian if this is seen in bivariate plots. See Brechmann and Joe (2015) for data examples that follow these steps.

2 Description of Datasets

Concrete Concrete compressive strength is a nonlinear function of age and ingredients. In this dataset, $n = 1029$ concrete samples are collected from 17 different sources (Yeh, 1998). There are 9 variables recorded: the concrete compressive strength (MPa), age (days) and 7 ingredients (kg/m^3). In order to run the brute-force algorithm, we only keep the age and ingredient variables, which gives a dataset with 8 variables.

Abalone The abalone dataset is obtained from the UCI machine learning repository (Lichman, 2013). It contains $n = 4177$ samples and 8 numerical variables, including age and physical measurements of abalones. It is also feasible to run the brute-force algorithm on this dataset.

Glioblastoma Tumors (GBM) The glioblastoma tumors dataset is a level-3 gene expression dataset studied by Brennan et al. (2013). It is obtained from The Cancer Genome Atlas (TCGA) Data Portal (Tomczak et al., 2015) and contains expression data of 12044 genes from $n = 558$ tumors. Within all the

genes in the dataset, we first filter out 1342 genes that are related to human cell cycle. Afterwards, a hierarchical clustering algorithm with Euclidean distance metric and complete-linkage is applied to obtain a cluster of 92 genes. To further study different scenarios, we randomly sample $d = 8, 10, 15, 20$ variables and repeat the procedure 100 times. This allows us to calculate confidence intervals when comparing different methods.

Deutscher Aktien Index (DAX) This dataset contains $n = 511$ daily log returns of 29 stocks listed in Deutscher Aktien Index (DAX) in 2011–2012 (Section 7.8.2 in Joe (2014)). A GARCH filter is applied to remove serial dependence. Similar to the sub-sampling procedure for the GBM dataset, we also randomly sample $d = 8, 10, 15, 20$ variables for 100 times.

References

Brechmann, E. C. and Joe, H. (2015). Truncation of vine copulas using fit indices. *Journal of Multivariate Analysis*, 138:19–33.

Brennan, C. W., Verhaak, R. G. W., McKenna, A., Campos, B., Noushmehr, H., Salama, S. R., Zheng, S., Chakravarty, D., Sanborn, J. Z., Berman, S. H., et al. (2013). The somatic genomic landscape of glioblastoma. *Cell*, 155(2):462–477.

Joe, H. (2014). *Dependence Modeling with Copulas*. Chpaman & Hall / CRC Press, Boca Raton, FL.

Kurowicka, D. and Cooke, R. (2003). A parameterization of positive definite matrices in terms of partial correlation vines. *Linear Algebra and its Applications*, 372:225–251.

Kurowicka, D. and Cooke, R. M. (2006). *Uncertainty Analysis with High Dimensional Dependence Modelling*. Wiley, Chichester.

Lichman, M. (2013). UCI machine learning repository.

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology*, 19(1A):A68.

Yeh, I.-C. (1998). Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*, 28(12):1797–1808.